A Proposal for using Metadata Encoding Techniques for Health Care Information Indexing on the WWW

Richard J. Appleyard, Ph.D., and Gary Malet, D.O.
Biomedical Information Communication Center,
Oregon Health Sciences University, Portland, OR, USA

The WWW is rapidly becoming a victim of its own success. Since there are currently no structured indexing features built into the WWW protocol, most of the efforts to improve health information retrieval, and information retrieval in general, have focused on independent indexing methods. However, the hand (manual) or computer (automatic) techniques currently available have significant drawbacks.

The sheer volume of information that is currently on-line necessitates automated indexing if a significant proportion is to be covered. But there is also an increasing amount of medical information that cannot be indexed externally because it is either not directly addressable or not obtainable without a password. The Harvest project [1] was developed to address the problems of centralized WWW indexing, such as scalability and network resource usage, using a distributed indexing model with search engines sharing indices. However, the common index format used could also be generated by password-protected sites and WWW-enabled databases. This would enable them to publicize what information they have available without compromising any proprietary data.

However, computer search engines still have difficulty identifying characteristics of media, such as the overall theme or type of a text document and any descriptive information about non-text media, e.g., pictures and video. This reduces the effectiveness of indexing the WWW, and particularly medical content that is comprised of rich, multimedia data. Manually-organized medical catalogs or WWW sites provide a limited answer to this by organizing medical information in a context-specific fashion. But the logistics of manual indexing make it unfeasible to cover more than the smallest fraction of the available information on the WWW.

One method to address this problem is to provide exactly what is missing from the documents - descriptive content or metadata [2]. The most detailed efforts to date are from a number of OCLC On-line Computer Library Center Metadata Workshops [3] and from the WWW Consortium's PICS (Platform for Internet Content Selection) [4] working group. Both groups have defined metadata structures and created methods for incorporating them into WWW pages. The "Dublin Core" is a set of meta-tags that define a base set of attributes about a WWW document, e.g., title, author, publisher, etc.

Obviously, different metadata sets will be needed to represent the different domains of expertise, such as Medicine. Authors would then need to include medical metadata tags within their WWW documents when publishing their multimedia content. However, this does not preclude author-independent assignment and publishing of these meta-tags, e.g., by current medical WWW indexing efforts, and thus a method of reviewing the quality of information.

However, in addition to medical descriptive content, metadata could be used to help solve other issues, such as the economics of information distribution. This is needed if the WWW is to evolve beyond being a medium for free or advertiser-supported information and facilitate access to proprietary information that is furnished for a fee. Meta-tags could describe the cost and payment details for the information.

These structured data elements and a common index format will allow WWW robots, such as Harvest (with some modifications), and WWW-enabled databases to create meta-indices of medical information on the WWW. This will allow the intelligent sorting of documents and more specific matching of Web pages to the information needs of users. Clinical medicine documents will be sortable by identifiers such as image formats and content, geographic markers, peer review status or specialty domain. The incorporation of nomenclatures, such as UMLS and SNOMED, into meta-tags will also allow the linking of Web-based knowledge sources into electronic medical record systems.

This poster will present possible metadata sets that could be used to describe medical content. It is hoped that this will provoke discussion about appropriate formats and the involvement of other interested groups in the development process. It is important that these meta-tag elements be defined in a rigorous way, and that a disciplined and collaborative framework be developed for the assignment of meta-tags. Their adoption by medical content authors will be crucial in bringing some order out of the current chaos on the WWW.

1. Bowman CM et al. The Harvest Information Discovery and Access System. Paper presented at the 2nd International WWW Conference, Chicago 1994 (http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/schwartz.harvest/schwartz.harvest.html).
2. Lynch C. Searching the Internet. Scientific American 1997(March).
3. The Dublin Metadata Workshop (Dublin, Ohio) March 1995, and The Warwick Metadata Workshop (Warwick, UK) April 1996.
4. Resnick P, Miller J. PICS: Internet Access Controls Without Censorship. Communications of the ACM 1996;39(10):87-93.

1091—8280/97/$5.00